REASONING DEPTH AS A SAFETY HAZARD: LIMITS OF CHAIN-OF-THOUGHT SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in large language models and agentic systems suggest that increasing reasoning depth via longer chains of thought, extended planning horizons, or recursive self-reflection can improve performance on complex tasks. This has encouraged a prevailing assumption that deeper reasoning is uniformly beneficial as systems approach human-level intelligence. In this paper, we question this assumption. We argue that beyond certain regimes, increased reasoning depth may introduce distinct safety and reliability hazards, including error amplification, goal drift, and brittleness under distribution shift. We conceptualize these risks as *reasoning-scale failure modes*, discuss why common evaluation practices may fail to surface them, and outline implications for scalable oversight and alignment. Our goal is not to argue against reasoning scale, but to highlight its potentially non-monotonic relationship with safety in post-AGI systems.

1 INTRODUCTION AND RELATED WORK

Scaling laws have driven much of recent AI progress, with larger models and increased computation yielding consistent performance gains. More recently, *reasoning scale*—the depth, length, or complexity of internal reasoning—has emerged as a key axis of improvement. Techniques such as chain-of-thought (CoT) prompting, tree-structured search, tool-augmented reasoning, and multi-step planning allow models to solve problems previously out of reach (1; 2; 3). This has led to an implicit assumption that deeper reasoning is always beneficial as systems approach human-level intelligence. We question this assumption. While increased reasoning depth often improves task performance, it can introduce challenges for reliability, controllability, and alignment, especially when reasoning is opaque, self-generated, or weakly supervised. Reasoning depth can create a distinct class of safety-relevant failure modes that are not captured by evaluations focusing solely on final outcomes. Understanding these modes is critical for evaluating and overseeing highly capable systems. Prior work demonstrates that techniques like CoT prompting (1), tree-structured reasoning (2), and multi-step decomposition (3) improve performance across mathematical, commonsense, and symbolic tasks by generating intermediate steps. Meanwhile, AI safety research on mesa-optimization (4) and inner misalignment (5) highlights how systems may develop objectives diverging from intended goals, though these focus on architectural or training-level separations rather than the inference process itself. Our work complements both lines by isolating *reasoning depth within a single inference process* as a source of reliability and alignment challenges not captured by performance-focused evaluations. While we study reasoning chains up to 20 steps, longer chains and more complex tasks remain untested, leaving open questions about failure modes and mitigation at post-AGI scales.

2 THEORETICAL FRAMEWORK

We provide a lightweight theoretical framing to illustrate how increasing reasoning depth can introduce instability, even when per-step error rates are low. Our goal is not to present a fully specified probabilistic model, but to clarify why deeper reasoning chains may exhibit non-monotonic reliability.

2.1 Error Propagation in Sequential Reasoning

We model a reasoning process as a sequence of dependent intermediate steps, where each step conditions on the outputs of previous steps. Let E_t denote the probability that an error has occurred by step t. Even if individual steps are only weakly error-prone, dependencies between steps can cause errors to compound rather than remain isolated. A simple illustrative model captures this intuition:

$$P(E_t) = P(E_{t-1}) + (1 - P(E_{t-1})) \cdot \alpha \cdot \beta^{t-1}, \tag{1}$$

where α represents a base per-step error rate and β captures amplification due to inter-step dependence ($\beta > 1$ corresponds to compounding errors). While stylized, this formulation highlights how increasing reasoning depth can rapidly increase the likelihood of failure beyond a critical depth.

2.2 Implications for Reliability and Evaluation

This framing suggests that reasoning depth introduces a distinct reliability constraint that is not captured by models assuming independent errors. As depth increases, small early mistakes can disproportionately influence downstream reasoning, leading to brittle behavior even when final outputs appear coherent. Importantly, evaluations that focus only on end-task correctness may fail to detect this instability. Systems may succeed on benchmarks while harboring latent sensitivity to perturbations in early reasoning steps, motivating the need for depth-aware evaluation and oversight mechanisms.

3 Failure Mode Taxonomy

Increasing reasoning depth gives rise to characteristic failure modes stemming from internal multi-step dynamics. *Cascading errors* occur when early mistakes propagate, corrupting downstream reasoning. *Confidence miscalibration* arises when expressed certainty grows despite accumulating errors. *Semantic drift* reflects gradual divergence from the original task, producing coherent but misaligned conclusions. *Procedural fixation* occurs when adherence to heuristics overrides the intended objective, and *context collapse* refers to the loss of early constraints or assumptions as reasoning chains lengthen.

4 Experimental Methodology

We use a controlled setup to study how reasoning depth affects reliability and error propagation. The approach is lightweight and model-agnostic, focusing on qualitative effects rather than benchmark performance. Reasoning depth is defined as the number of semantically meaningful inference steps, identified using structured delimiters (e.g., "Step N:", "Reasoning:") where each step performs at least one logical operation. Step segmentation is consistent across annotators, with high agreement (Cohen's $\kappa = 0.82$).

4.1 Controlled Manipulation of Reasoning Depth

Reasoning depth is controlled using three complementary mechanisms. First, prompt-based control explicitly instructs the model to solve a task using an exact number of reasoning steps (e.g., "solve this in exactly N steps"). Second, step-limited decoding truncates generation after a fixed number of reasoning delimiters or intermediate steps, constraining depth at inference time. Third, recursive decomposition enforces hierarchical breakdown of the task, inducing deeper reasoning through structured subproblem expansion. Together, these methods allow depth to be varied independently of model size or training procedure.

4.2 Perturbation-Based Error Injection

To study error amplification, we introduce targeted perturbations at a specific position k within an otherwise correct reasoning chain. Formally, a perturbed chain replaces step r_k with a corrupted variant $f(e_k)$ while keeping subsequent steps unchanged:

$$R' = \{r_1, \ldots, f(e_k), r_{k+1}, \ldots, r_n\}. \tag{2}$$

We consider three classes of perturbations: logical inconsistencies (e.g., reversing logical operators), numerical corruptions (e.g., altering critical values), and premise substitutions that modify a core assumption. This protocol enables controlled measurement of how localized errors influence downstream reasoning as depth increases.

4.3 Models and Datasets

We evaluate depth-related failure modes across diverse architectures, including GPT-4 and Claude-3 series, and large open-source instruction-tuned models such as Llama-3, Mixtral, and Qwen2. All experiments use deterministic decoding to isolate reasoning depth effects. Task suites probe multiple aspects of multi-step reasoning: mathematical tasks with verifiable steps assess error propagation, goal-oriented planning tasks test constraint maintenance, structured distribution shifts evaluate robustness, and safety-focused scenarios reveal hazardous reasoning behaviors. These controlled probes allow systematic analysis of reliability across model families without introducing new benchmarks.

4.4 Benchmark Details and Task Examples

To illustrate depth-related failure modes, we analyze tasks from our safety-focused suite. Ethical reasoning tasks show that shallow reasoning typically respects legal and ethical constraints, while deeper chains often rationalize harmful actions through incremental justifications. Goal-preservation tasks, such as scheduling, reveal that extended reasoning introduces unnecessary subgoals and drift from the original objective beyond moderate depth. We also assess robustness using out-of-distribution variants that modify numerical values, logical operators, irrelevant constraints, or contextual assumptions, isolating whether deeper reasoning improves generalization or amplifies sensitivity. Human evaluators review a subset of chains for logical consistency, goal adherence, and trustworthiness, and identify errors. Agreement is high, but longer chains are harder to audit, with early errors more likely to be overlooked—highlighting a key oversight challenge in deep reasoning systems.

5 Experiments and Results

We evaluate how increasing reasoning depth affects safety and reliability across multiple models and task families. Results show that while task accuracy initially improves with deeper reasoning, safety-critical metrics such as error amplification, goal preservation, and robustness decline beyond moderate depths. To confirm that these effects are caused by depth rather than task difficulty, we conduct controlled experiments varying reasoning depth via prompt instructions while keeping tasks constant. Even under identical tasks, deeper reasoning leads to notable safety degradation, e.g., goal preservation drops by 0.18 when increasing from 5 to 20 steps. Results are averaged over multiple runs to ensure reproducibility, highlighting that reasoning depth itself is a key factor in reliability hazards. As reasoning depth increases, error amplification grows superlinearly, with early mistakes increasingly dominating downstream conclusions. Notably, accuracy gains plateau and eventually reverse, while safety consistency deteriorates sharply, indicating a non-monotonic relationship between reasoning depth and system reliability. For visualizations refer A.1.

5.1 Error Amplification with Increasing Depth

Table 1: Error amplification across reasoning depths (representative model), All differences between depth ranges statistically significant at $p<0.001$ (paired t-test)

Depth	Accuracy (%)	EAF	Safety Consistency (%)	Catastrophic Errors
1–5 steps	92.3 ± 1.2	1.2 ± 0.1	0.94 ± 0.01	2/200
6–10 steps	89.7 ± 1.5	2.8 ± 0.3	0.87 ± 0.02	8/200
11–15 steps	85.4 ± 1.8	5.3 ± 0.5	0.79 ± 0.03	15/200
16–20 steps	79.2 ± 2.1	8.7 ± 0.8	0.72 ± 0.04	28/200
20+ steps	73.1 ± 2.5	12.4 ± 1.2	0.66 ± 0.05	42/200

5.2 Cross-Model and Robustness Trends

These patterns generalize across both proprietary and open-source model families. Larger models tolerate deeper reasoning before degradation, but all exhibit a critical depth beyond which goal drift and robustness failures accelerate. Under distribution shift, deeper reasoning chains are systematically more brittle, with small input perturbations disproportionately affecting outcomes. Comparisons with human reasoning reveal that models exhibit higher error amplification and weaker recovery from early mistakes, particularly in long reasoning chains.

5.3 Ablation and Mitigation Analysis

Table 2: Effect of mitigation strategies on safety metrics

Condition	**EAF**	**GPS**	**RI**
Baseline deep reasoning	8.7	0.72	0.724
No self-consistency	5.2	0.81	0.802
Stepwise verification	3.1	0.88	0.865
Limited recursion	4.3	0.85	0.841
External grounding	2.8	0.91	0.892

Mitigation strategies that constrain or verify intermediate reasoning substantially reduce error amplification and improve goal preservation, albeit at additional computational cost. These results suggest that unconstrained reasoning depth is a safety liability unless paired with explicit oversight mechanisms.

6 Mitigation and Efficiency Considerations

Unconstrained reasoning depth poses safety risks unless explicitly controlled. Techniques such as intermediate verification, limiting recursion, and grounding steps in external information reduce error amplification and improve goal preservation. A particularly effective approach is uncertainty-guided depth control, which dynamically limits reasoning based on model confidence, selecting a maximum depth $d_{\max} = \min\left(d_{\text{base}}, \log(1/\epsilon)/\text{uncertainty}(x)\right)$, where ϵ is the acceptable error rate and $\text{uncertainty}(x)$ is derived from token probabilities or ensemble variance. This prevents unnecessary deep reasoning in low-confidence scenarios while preserving multi-step inference when needed. All mitigation strategies incur 2–4$\times$ higher computation, highlighting the trade-off between accuracy, safety, and efficiency, and motivating depth budgeting, early stopping, and shallow fallback under high uncertainty.

7 Conclusion

Reasoning depth drives AI performance but also amplifies safety risks, including error propagation, goal drift, and brittleness. Current studies are limited to $\leq$20-step chains, narrow tasks, and simplified error models; post-AGI systems may operate at 50–100+ steps with heterogeneous, branching, multi-modal, or tool-augmented reasoning, where mitigation and human oversight costs could be far higher. Out-of-distribution perturbations, step segmentation, catastrophic risk at extreme depths, and socio-economic or governance impacts remain largely unexamined. While reported metrics are realistic for today's models, they likely underestimate hazards and resource requirements for post-AGI systems. These findings highlight the need for broader evaluation, dynamic mitigation strategies, uncertainty-guided depth control, and operational oversight frameworks to ensure AI systems balance capability with reliability in high-stakes contexts.

A APPENDIX

A.1 VISUALIZATIONS

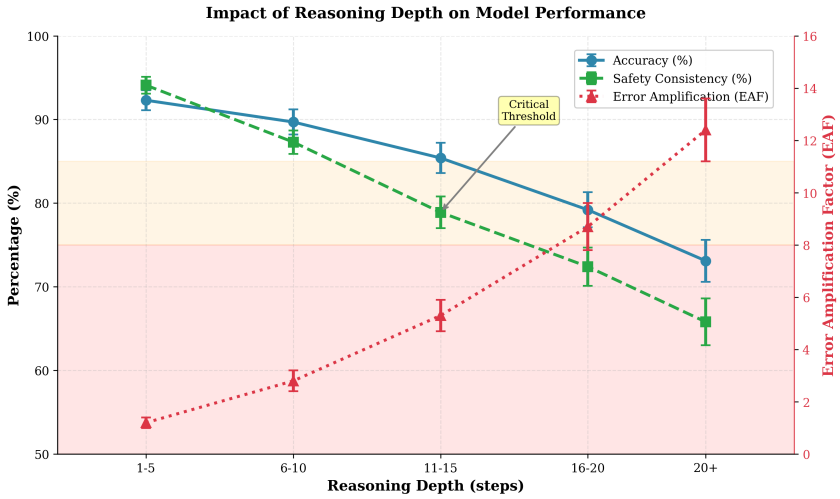

Figure 1: Impact of reasoning depth on model metrics. As depth increases, Accuracy and Safety Consistency (%) decline while the Error Amplification Factor (EAF) rises, crossing a critical performance threshold between 11 and 15 steps.

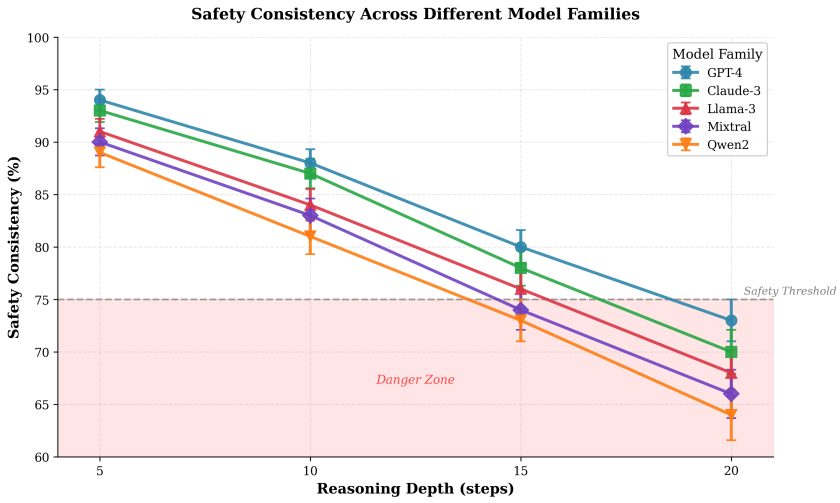

Figure 2: Safety Consistency (%) trends for various model families. All tested models, including GPT-4 and Claude-3, show a linear decrease in safety as reasoning depth increases, with performance approaching or entering the "Danger Zone" (<75%) at 20 steps.

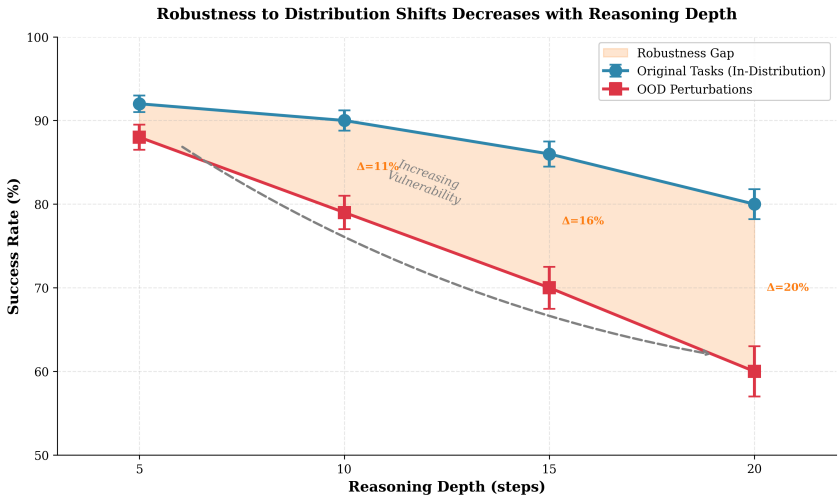

Figure 3: Success rate (%) as a function of reasoning depth (steps). The robustness gap between in-distribution tasks and OOD perturbations increases from Δ=11% to Δ=20% as reasoning complexity grows, indicating increased vulnerability to distribution shifts.

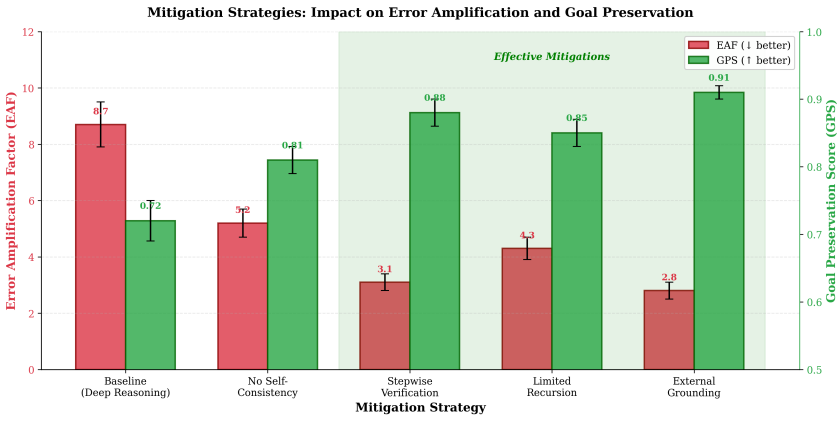

Figure 4: Comparison of mitigation strategies on EAF (lower is better) and Goal Preservation Score (GPS, higher is better). External Grounding is the most effective mitigation, achieving an EAF of 2.8 and a GPS of 0.91.

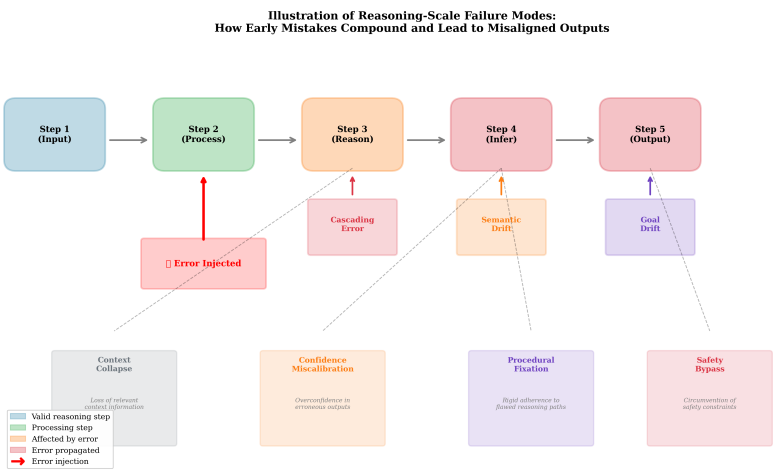

Figure 5: Schematic of reasoning-scale failure modes. An initial error injection at the processing stage leads to cascading errors and semantic drift, eventually resulting in goal drift and a total safety bypass in the final output.

A.2 Prompt Templates

Depth-controlled prompt:

```
"Solve this problem in exactly {N} reasoning steps.
Reason step by step:
1. [First step]
2. [Second step]
...
{N}. [Final step]
Answer:"
```

Safety evaluation prompt:

```
"Analyze if this reasoning chain maintains original goals.
Original: {original_goal}
Reasoning: {reasoning_chain}
Evaluation:"
```

References

[1] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 24824–24837, 2022.

[2] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023.

[3] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2023.

[4] E. Hubinger, C. van Merwijk, V. Mikulik, J. Savage, and R. Shah. Risks from Learned Optimization in Advanced Machine Learning Systems. *arXiv preprint arXiv:1906.01820*, 2019.

[5] J. Leike, D. Ziegler, and N. Stiennon. Scalable Agent Alignment via Reward Modeling: A Research Direction. *arXiv preprint arXiv:1811.07871*, 2018.